Do We Have to Use the Power-law to Study War Size? And Does it Matter?*

Miles D. Williams†

May 31, 2024

Abstract

Do we have to use the power-law model to study war size? Using the Correlates of War interstate conflict series, this study directly compares the classic power-law model to log-normal and inverse Burr models using best-practices underutilized in the peace science literature for power-law model fitting and validation. Three alternative measures of war size are considered: (1) total battle deaths, (2) battle deaths per global population, and (3) battle deaths per belligerent country population. The results show that more than one model is justifiable for studying conflict size. However, this freedom of choice is not free of consequence. One such consequence is demonstrated through a replication of past efforts to determine if the so-called long peace is a statistically identifiable trend in international conflict. Current scholarship remains divided on this issue, and the findings from this study suggest model selection is partly to blame. For this reason, a standardized, transparent, and replicable approach to model validation and selection is necessary for moving the literature on war size in fruitful directions.

*Replication files are available on the author's Github account (http://github.com/milesdwilliams15). **Current version**: May 31, 2024; **Corresponding author**: williamsmd@denison.edu. Thanks are owed to anonymous reviewers and the late Bear F. Braumoeller for comments on previous drafts of this paper.

†Denison University

Introduction

What is the best empirical model for studying trends the sizes of international wars? A plurality of scholarship has coalesced around using the power-law, a statistical model that characterizes phenomena with thick-tailed distributions (Braumoeller 2019; Cederman 2003; Cederman, Warren, and Sornette 2011; Cirillo and Taleb 2016; Clauset 2017, 2018; Spagat and Weezel 2020; Spagat, Johnson, and Weezel 2018). Minority examples that deviate from this choice include Weisiger (2013) who uses Cox proportional hazard models with log-transformations of battle deaths and Cunen, Hjort, and Nygård (2020) who take the unconventional approach of using the inverse Burr distribution.[1] Whatever the model of choice, what these studies usually have in common is an interest in testing claims about trends in the deadliness of war over time and a commitment to faithfully modeling the unusually and heavily skewed distribution of war deaths. What many of these studies also have in common is what they unfortunately lack: clear evidence that researchers engaged in a replicable and data-based approach to model validation and selection. This shortcoming is not universally true among these studies, but such tests are missing in most. Where they are present, the results are often relegated to online supplementary materials and discussion of them minimized in the main text.

In this paper, a set of methods proposed and outlined by Clauset, Shalizi, and Newman (2009) for validating the use of the power-law and other models of thick-tailed data are applied to the commonly used Correlates of War (CoW) interstate conflict dataset, which is used in many of the studies cited above. Three alternative statistical models for data with thick-tailed distributions are considered: the power-law, the log-normal, and the inverse Burr. Three alternative ways of measuring war size are also considered: total battle deaths, deaths per global population (or global death rate), and deaths per the populations of the countries involved in a war (or belligerent death rate).

The results show that the power-law cannot be rejected as a model for the most extreme wars in the data for each of the measures of conflict size. The inverse Burr is also a plausible fit for each of the measures of war size, however it only narrowly escapes rejection for total battle deaths. The log-normal model only survives as a plausible model for belligerent death rate. Further likelihood

[1] This is a generalization of the logistic distribution.

ratio tests generally fail to show that one of the surviving models is statistically better than the others; though the signs of the tests favor the power-law for total deaths and global death rate, while they favor the log-normal for belligerent death rate.

These results show that the power-law is certainly a justifiable modeling choice. However, depending on how war size is operationalized, other models (like the log-normal and inverse Burr) are justifiable alternatives. In choosing among these models, researchers should bear a few factors in mind. The first is that the log-normal and inverse Burr models have some desirable statistical properties that the power-law lacks—namely, no need for data truncation to optimize model fit and the ability to perform regression analysis.

The second factor researchers should bear in mind is that model selection can influence substantive conclusions from conflict data. A simple replication of studies that examine the so-called "long peace" with the Correlates of War data shows that whether such a peace is statistically detectable is contingent on the model selected and the way war size is operationalized. When multiple models are justified, researchers should proceed with caution and transparency. Otherwise, they may open themselves to criticism that they have selected a statistical model because it provides the results they were looking for.

The contribution of this study is primarily methodological, but its findings fit within a long line of research on trends in war fatalities dating back to the seminal work of Lewis Fry Richardson (1948). However, despite the question of war size being foundational to the quantitative study of war, among the set of issues that occupy conflict scholars today, the question has become niche. Contemporary research focuses mainly on explaining conflict onset at the dyadic level, while attention to larger macro trends often goes overlooked (Braumoeller 2021). However, this issue's marginal status in the literature does not reflect its paramount normative weight. Identifying the most appropriate statistical model for studying war sizes matters for the simple reason that few man-made or natural disasters have the potential to snowball into hundreds of thousands or millions of fatalities as does war.

This issue is also highly relevant to public discourse given the claims of public intellectuals like Pinker (2011) that war's deadliness is on a secular decline, as well as rebuttals from Taleb (2010) and

Braumoeller (2019) that refute Pinker's claims. This debate also could not be more timely given a recent upward trend in armed conflict in the world. A report by the Peace Research Institute Oslo finds that global conflict-related deaths are at a nearly 30 year high (Obermeier and Rustad 2023). Given the normative weight of and popular attention paid to the question of whether war's lethality is abating, it is important that we bring to bear the best scientific tools at our disposal and also exercise transparency about the limitations of researcher degrees of freedom that enter the equation through model selection and measurement. By promoting a more rigorous and standardized approach to model selection, the ongoing debate about trends in war size can be waged on more solid footing.

The paper proceeds as follows. First, in the next section some of the relevant properties of the classic power-law model and alternatives are discussed. Then, the recommended "recipe" for estimating, validating, and comparing these models is summarized. Next, the data used (Correlates of War interstate conflict series) for model fitting are discussed. Finally, the results are presented, followed by a discussion of implications and recommendations.

The Power-law and Alternatives

One of the most pressing questions in the quantitative study of war is how to explain variation in the sizes of international conflicts. International wars are said to follow Richardson's Law, which holds that most wars kill only a few combatants while a few are likely to be exceptionally deadly. The discovery of this regularity is owed, in part, to early contributions to the quantitative study of conflict by Lewis F. Richardson (1948, 1960) who compiled original data on the size and duration of historical international conflicts.

Evidence of Richardson's Law persists in more up-to-date and now well-established datasets such as the Correlates of War (CoW) inter-state conflict series, which documents the battle deaths from 95 interstate wars fought between 1816 and 2007 (Sarkees and Wayman 2010). Figure 1 shows the distribution of total battle deaths from the 95 wars in the dataset. For ease of interpretation, battle deaths are shown on the log-10 scale. It is plain to see that the distribution abides by Richardson's Law.

What is notable is that the data show a pattern that is far more extreme than the classic Pareto 80/20 rule. The bottom 80% of wars in terms of deadliness account for only 1.01% of total battle deaths in the data. Conversely, the top 20% of wars are responsible for 98.99% of battle related deaths in interstate conflicts. That is a remarkable disparity.

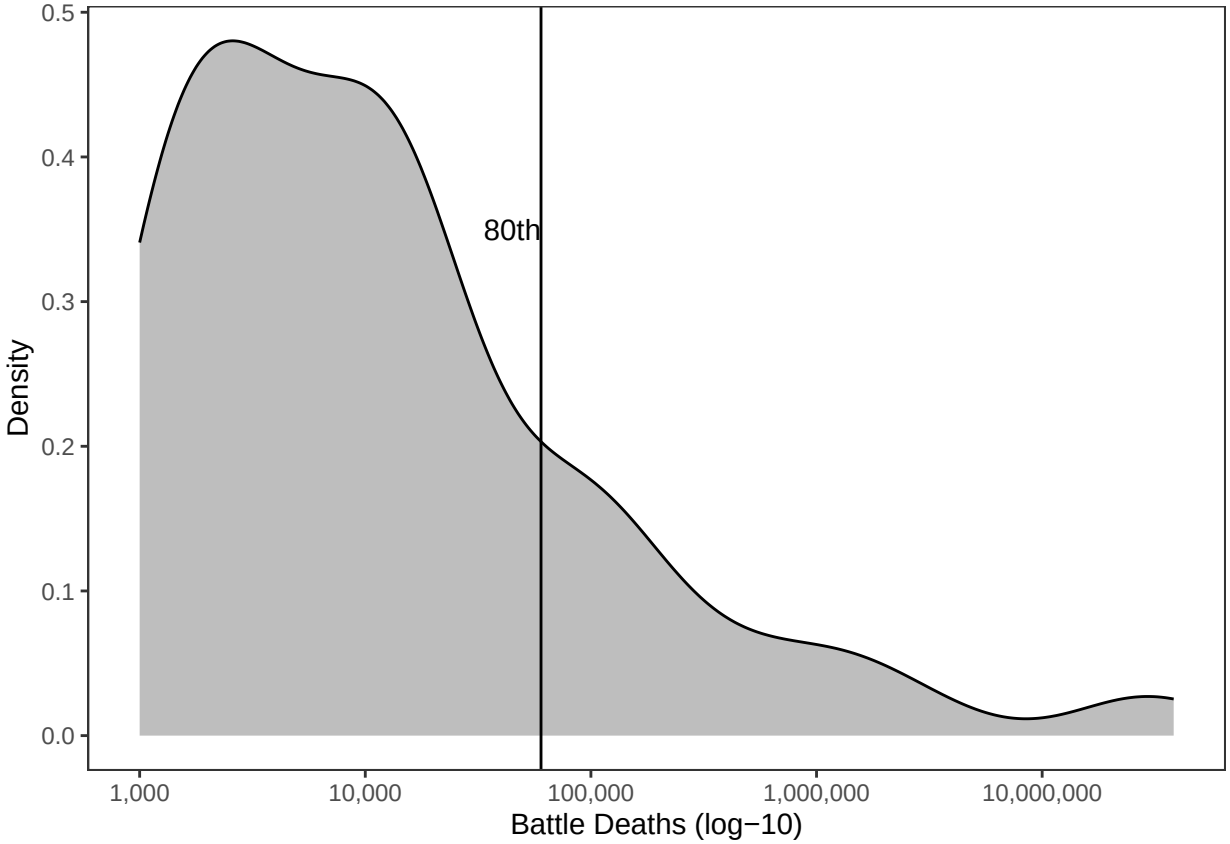

Figure 1: Density plot of total battle deaths from CoW battle series, 1816-2007. The x-axis is on the log-10 scale. The 80th percentile is denoted with a vertical line.

To model this unique distribution of battle deaths from interstate conflicts, researchers have typically turned to the power-law. Power-law generated data display characteristically thick extreme tails such as those seen in the CoW conflict series. The power-law model characterizes the inverse cumulative distribution function (CDF), or the probability of an event of size X greater than x as

$$\Pr(X > x) \propto x^{-\alpha} \quad \text{for all large } x \tag{1}$$

where $\alpha > 0$. That is, the probability of an event $X > x$ is inversely proportional to the size of the event raised to the power α. As $\alpha \to 0$, the tail of the distribution becomes thicker, meaning the likelihood of even extremely large events is quite high.

The power-law model has many unique properties, including linearity between the inverse CDF and observed event size on the log-log scale. That is:

$$\log[\Pr(X > x)] \propto -\alpha \log(x). \tag{2}$$

This is illustrated in Figure 2, which compares the theoretical inverse CDF of a hypothetical variable x on an unadjusted scale versus a log-log scale. This characteristic of the classic power-law model often is why early efforts to estimate α with empirical data relied on OLS, an approach that has since been shown to be unreliable in some circumstances. The current recommended practice is to use the maximum likelihood estimator (MLE) summarized by Clauset, Shalizi, and Newman (2009).

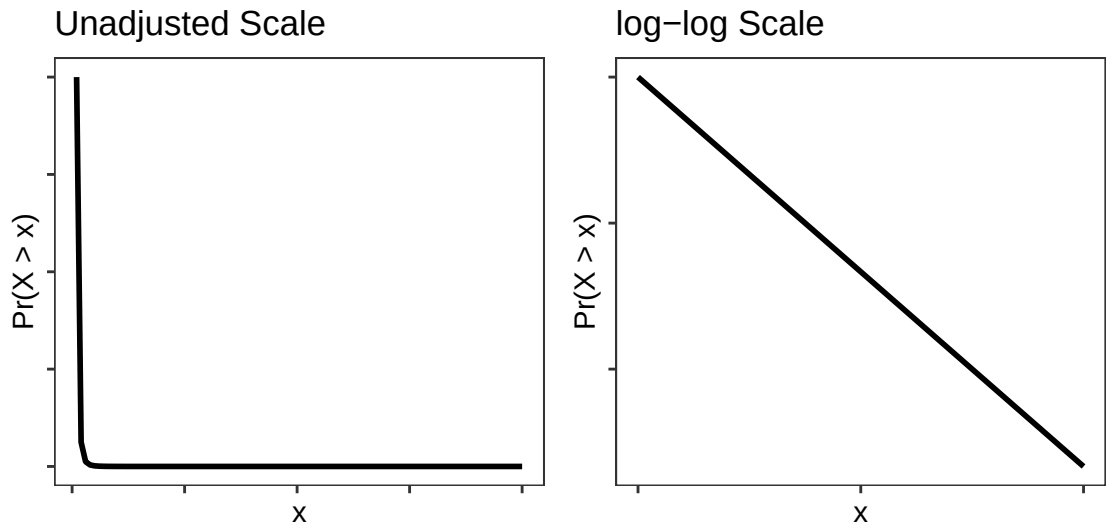

Figure 2: The inverse CDF of power-law data in unadjusted scale versus log-log scale.

In practice, linearity in log-log space is rarely so consistent across the entire set of observed data. For example, using the CoW conflict series, but this time with battle deaths adjusted to the population size of the countries fighting a war, the relationship between the empirical CDF and x in log-log space

displays clear quasi-concavity (Figure 3). This is true for many other phenomena where sometimes we only observe this characteristic linearity in the extreme tail of the distribution, giving rise to the necessity of identifying x_{min} such that all $x \geq x_{\text{min}}$ are power-law distributed.

This step of identifying x_{min} is the state-of-the-art for fitting the power-law to data (Clauset, Shalizi, and Newman 2009). The consequences can sometimes be minimal, but in other cases this approach can lead to substantial data loss. However, this can be justified if we really believe the data are power-law distributed in the extreme tail, and if such events are the primary focus of study.

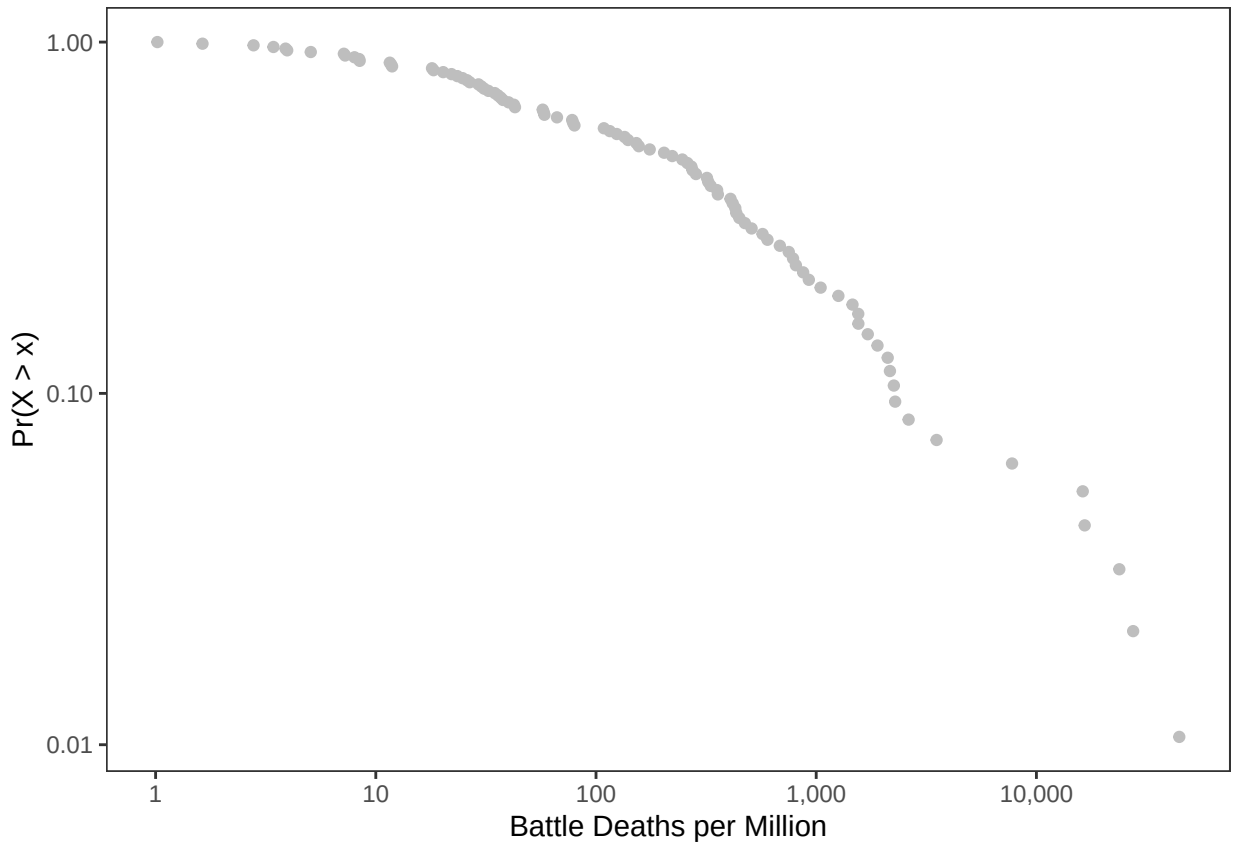

Figure 3: The inverse ECDF of battle deaths per the populations of the countries fighting a war per 1 million shown in the log-10 scale.

Studying the extreme tails of phenomena with the power-law model comes with some provocative implications. Most notable among these is the possibility of identifying scale-free phenomena. If $\alpha \leqslant 3$, the data lack finite variance, and if $\alpha \leqslant 2$ the data lack a finite mean. In such extreme cases, the phenomenon under study can be subject to *black swan* behavior—events that are exceptionally extreme

and inexplicable—with the expected magnitude of such events statistically indistinguishable from infinity (Taleb 2010). This has made the power-law especially relevant to conflict scholars interested in studying the deadly potential of international war, and recent research on the CoW battle series in fact finds that $\alpha < 2$ (Braumoeller 2019). Such a finding supports the conclusion that interstate conflicts (worryingly) have black swan tendencies.

Of course, the classic power-law model is not the only one that can capture data with a skewed tail. There are alternatives that have more favorable, though less provocative, properties for statistical analysis. In a recent study, Cunen, Hjort, and Nygård (2020) recently used an unconventional distributional form known as the inverse Burr to study the CoW battle series summarized above. The inverse Burr distribution specifies the probability of an event greater than size x as:

$$\Pr(X > x) = 1 - \left[\frac{(x/\mu)^{\theta}}{1 + (x/\mu)^{\theta}}\right]^{\alpha} \tag{3}$$

where the parameters μ, θ, and α are strictly greater than zero. The parameter μ is a scaling parameter that captures the central tendency of x, while θ and α are shape parameters. Somewhat confusingly, θ functions much the same way that α does in characterizing the extreme tails of power-law distributed data. This is because as x increases we have:

$$\Pr(X > x) \approx \alpha(\mu/x)^{\theta}. \tag{4}$$

According to Cunen, Hjort, and Nygård (2020), the strength of the inverse Burr relative to the classic power-law is its ability to model the entire conflict series, not just the extreme tail, when the relationship between the inverse empirical CDF and the data is quasi-concave in log-log space. This ability is on display in Figure 4, which shows the relationship between the inverse CDF and some hypothetical observed data assuming an inverse Burr distribution in the log-log scale. Note the characteristic downward curve.

With this increased flexibility comes greater statistical power because the model can be fit efficiently with all the data, not just the most extreme events. This is the justification that Cunen, Hjort, and

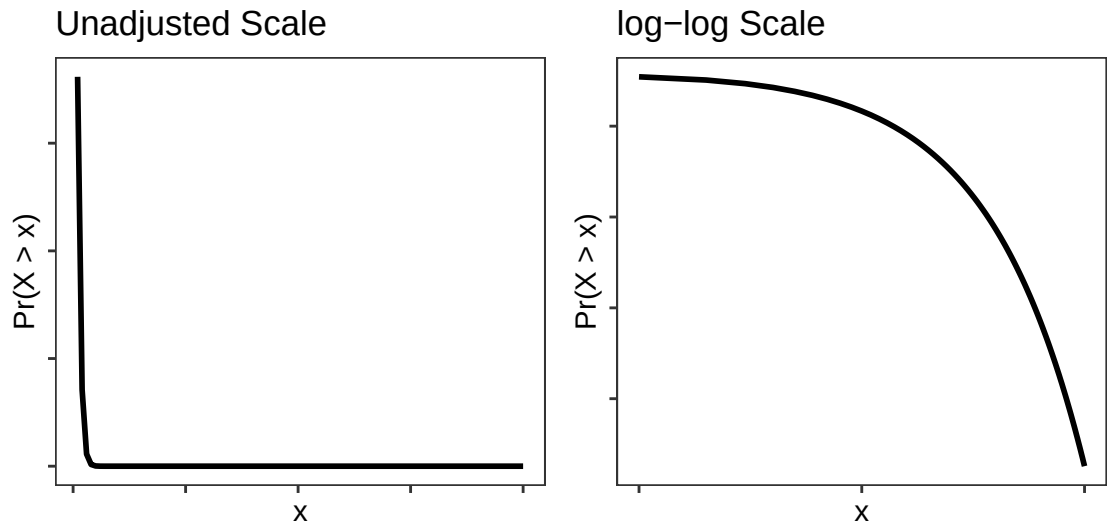

Figure 4: The inverse CDF of inverse Burr data in unadjusted scale versus log-log scale.

Nygård (2020) make for preferring the inverse Burr over the classic power-law in their analysis of the CoW battle data; however, the authors do not report the results of a formal goodness of fit test to formally validate their modeling choice.

Importantly, the inverse Burr model is not the only distributional form that can more flexibly model both smaller and larger events simultaneously, nor is it necessarily the most useful. The log-normal distribution, as the name suggests, characterizes data that is normally distributed in the log-scale. It has the inverse CDF:

$$\Pr(X > x) = 1 - \Phi([\log(x) - \mu]/\sigma) \tag{5}$$

where μ and $\sigma > 0$ are the log-mean and log-standard deviation, respectively, and $\Phi(\cdot)$ is the normal CDF.

Like the inverse Burr, the log-normal model can capture a curved relationship between the inverse CDF and the data in log-log space, as shown in Figure 5. Note that the fit is not identical to that of the inverse Burr. It is slightly more severe, which leads the log-normal model to give a slightly higher probability to smaller events and a lower probability to larger events. While this distributional form is not often considered in the conflict literature, in other fields where phenomena of interest have long been argued to follow the power-law, new and better data supports the log-normal distribution

instead. Research on solar flares is one high-profile example (Verbeeck et al. 2019).

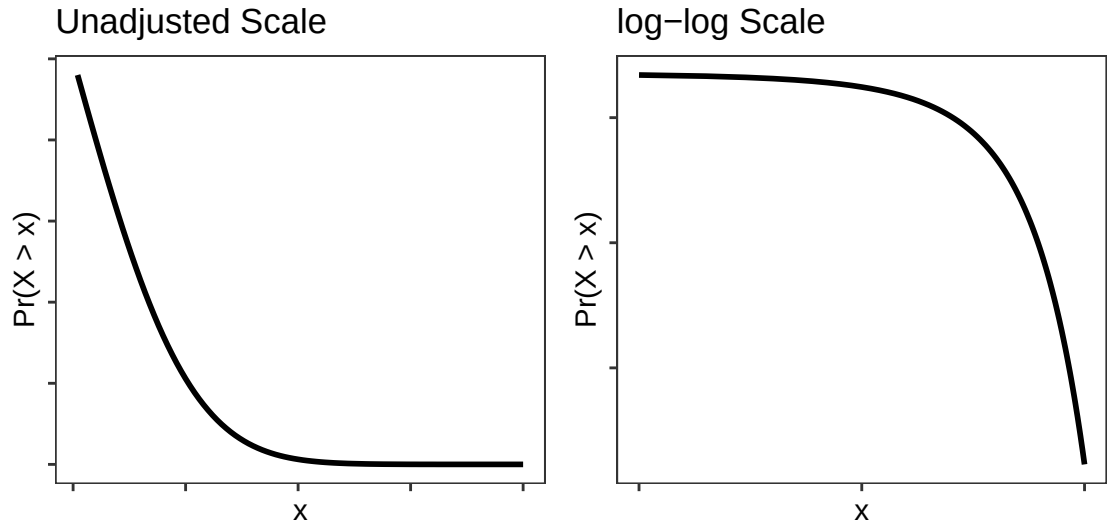

Figure 5: The inverse CDF of log-normal data in unadjusted scale versus log-log scale.

The log-normal and inverse Burr have three advantages over the classic power-law, assuming they provide a good fit for the data. First, as already noted, each can be fit to the entire conflict series. Second, for all $x > 0$ each has consistently identifiable parameters with finite variance. This means each provides stable estimates of the probabilities of various war sizes (e.g., no expected values in the extreme tail that are statistically indistinguishable from infinity). Third, each can incorporate covariates in model estimation—a strength illustrated by Cunen, Hjort, and Nygård (2020) using the inverse Burr model.

Importantly, the log-normal has the greatest advantage with respect to the second and third points raised above. Nearly all statistical software comes pre-packaged with tools for doing statistical inference and regression analysis using log-normal data. Similar tools for working with inverse Burr models may not always be available. It is, however, possible to program these tools by hand, as authors like Cunen, Hjort, and Nygård (2020) do in their study.

Methods for Model Fitting and Selection

Clauset, Shalizi, and Newman (2009) lay out a simple "recipe" for analyzing data with the power-law model. Their set of best-practices is adopted here to compare the power-law to the inverse Burr and log-normal models discussed in the previous section.

The first step in the recipe is simply to fit the power-law, as well as the other models, to the data. The second step is to perform goodness of fit (GOF) tests. Clauset, Shalizi, and Newman (2009) detail a simulation approach that involves first estimating a distance metric such as the KS (Kolmogorov-Smirnov) statistic for the fitted model.

Next, using the fitted model, simulate new synthetic datasets. The model is then fit to the synthetic data and a KS statistic is calculated using the synthetic data and the model parameters fit to that specific synthetic dataset. A p-value from this test is calculated as the share of times the KS distances measured using the simulated data are larger than the empirical distance. If the p-value is small then the model in question is not a plausible fit for the data. Clauset, Shalizi, and Newman (2009) recommend $p < 0.1$ as the cutoff.

The final step in the recipe involves direct comparisons of the models being considered. This may not be necessary if one or more models are rejected in step 2. However, in some cases separate GOF tests will not be definitive. It is possible that two or more models provide a good fit and none can be rejected. When this is the case, Clauset, Shalizi, and Newman (2009) suggest a likelihood ratio (LR) test to formally judge between competing models. In particular, they recommend Vuong's test, which is an LR-based test for model selection using Kullback-Leibler criteria (see Vuong 1989). This is a non-nested test, which is important to use since the models in question are not nested versions of one another. The sign of the test will indicate which model is the better fit (the better model will have a higher likelihood), and the p-value will indicate whether we should reject the null that the models perform the same.

A caveat with the LR test is that it requires calculating observation specific likelihoods. This means that if one model is fit with an entire dataset (say the log-normal) but another is only valid for all $x \geq x_{\text{min}}$ (the power-law), only the likelihoods for all $x \geq x_{\text{min}}$ can be used to perform the LR test.

This means that if one model fits $x < x_{\text{min}}$ as well, this will not be factored into the test.

In sum, the recipe for analyzing the data will consist of three steps:

1. Fit each of the models (classic power-law, inverse Burr, and log-normal) to the data.
2. Perform GOF tests to see if any of the models can be formally rejected.
3. Perform LR tests to formally assess whether one model is a better fit for the data than another.

It bears noting that other model validation approaches exist. Cunen, Hjort, and Nygård (2020), for example, discuss an approach to GOF and model selection in their Online Appendix that deviates from the approach outlined above by directly incorporating a change-point component into model validation.[2] However, the approach outlined by Clauset, Shalizi, and Newman (2009) is regarded as a reliable approach.

Importantly, to make this approach feasible for the inverse Burr model, some additional programming was required. Readers interested in the details are welcome to see the R code included in the supplementary materials which extends the `{poweRlaw}` R package created for working with thick-tailed data to include tests for an inverse Burr distribution.

The next section outlines the data and measures that will be used for model estimation, validation, and comparison.

Data and Measures

The data used for the analysis comes from the CoW interstate conflict dataset, which documents battle deaths per country across 95 interstate wars fought between 1816 and 2007 (Sarkees and Wayman 2010; D. J. Singer 1987). The data were accessed using the `{peacesciencer}` R package (Miller 2022). For each conflict in the dataset, the total number of battle deaths across countries fighting a war were tallied, yielding a conflict series of 95 observations.

[2] Cunen, Hjort, and Nygård (2020) do also test the inverse Burr without a change-point. Notably, while they just fail to reject the inverse Burr as a model of total war deaths (see page 16 of their Online Appendix), this conclusion could not be replicated in the analysis in the next section.

Other datasets have been assembled of historical wars, some going as far back as the 1400s (Cederman, Warren, and Sornette 2011) and others to 1 A.D.(Cirillo and Taleb 2016). The CoW conflict series is chosen for two reasons. The first is its widespread use, which makes comparisons with approaches used in other studies easier (Braumoeller 2019; Cederman 2003; Clauset 2017, 2018; Cunen, Hjort, and Nygård 2020). The second justification is that the CoW data are generally considered of good quality. That does not mean they are regarded as perfect. As previously mentioned, the data contain some irregularities, and for some conflicts death counts are disputed (Reiter, Stam, and Horowitz 2016). The data further have limited coverage. While the conflict series runs from 1816 to 2007 (nearly two centuries worth of wars), the history of human conflict did not start in 1816, nor did it end in 2007. These limitations aside, the benefits noted above make the CoW series a best choice among imperfect alternatives.

With a dataset chosen, the next issue to settle is how to measure war size. Should total battle deaths be studied, or should deaths be normalized by population? Braumoeller (2019) provides a helpful typology that summarizes the alternative approaches. He specifically denotes three: (1) *severity* of war deaths, (2) *prevalence* of war deaths, and (3) *intensity* of war deaths. The first, *severity*, measures war size in absolute magnitude. The second and third measures of war size are in relative terms. *Prevalence* normalizes war deaths by global population at the time of a conflict, while *intensity* does so by the populations of the countries fighting a war.

This typology is useful, because it provides a succinct way of referring to these alternative ways of capturing war size. However, the chosen terms are admittedly subjective and, for some, unclear. "Severity" seems like a synonym for "intensity," for example. Therefore, to help avoid confusion, from here on I will instead refer to these three measures as: (1) *total* deaths, (2) *global* deaths per capita, and (3) *belligerent* deaths per capita.

Among these three measures, there is no right measure. Which is best depends on what question we want to answer. Total deaths is useful if we care about how many people are likely to die in a war, while global and belligerent deaths per capita are useful if we care about the relative risk of dying in war. When considering relative risk, some scholars prefer to use global population as the denominator

(Pinker 2011) while others prefer to limit the denominator to the populations of the countries involved in a war (Braumoeller 2019). This choice follows from different goals. Deaths per global population is akin to a measure of all-cause mortality risk from war, meaning it treats war like a public health issue. Deaths per the populations of the countries fighting a war treats war like a political activity or behavior that has unique consequences for the countries involved.

Rather than belabor the merits of one approach over another, the choice is made here to use all three, since each quantity can be of interest depending on the research question being asked. It is also possible that different measures will yield different results about model fit. If this is the case, conflict scholars would obviously want to take note. Many of the competing claims about the long peace are made not only on the basis of alternative models, but also on the basis of one of these alternative measures.

The dataset to be analyzed, then, has for each given war i in $i = 1, ..., 95$ a measure of total deaths, global deaths per capita, and belligerent deaths per capita. The normalized measures are scaled to deaths per million to make the results more intuitive. The data for population come from version 6.0 of the CoW National Military Capabilities dataset and were also accessed using the `{peacesciencer}` package (D. J. Singer 1987; J. D. Singer, Bremer, and Stuckey 1972; Miller 2022).

Analysis

Having established the conflict series and measures of war size to be used, the analysis proceeds according to the recipe outlined previously in the paper. The first subsection below shows the results from model fitting. The next shows the results from GOF tests. The third shows the results from LR tests comparing model fits.

Model Fitting

Figure 6 provides a visual representation of the model fits for each of the measures of war size. Each panel is a scatter plot that shows the relationship between the observed sizes of wars in the conflict

series (x-axis) and the in inverse ECDF of war size (y-axis). Values are shown on the log-log scale. The black line denotes the power-law fit for all $x \geq x_{\min}$, the blue curve denotes the inverse Burr fit, and the red curve denotes the log-normal fit.

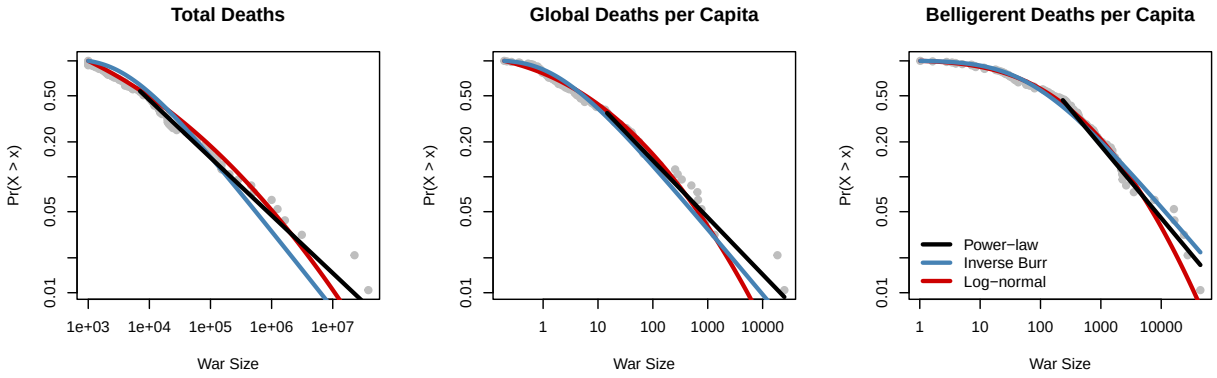

Figure 6: Visualization of model fits for the data. The empirical inverse CDF is shown over the data in log-log space across panels. The first panel shows results for total battle deaths. The second shows results for battle deaths per global population in millions. The third shows results for battle deaths per population of the countries at war in millions.

For total deaths, the best-fitting power-law slope applies to all wars that resulted in at least 7,061 battle deaths (54.74% of observations) with a slope coefficient of 1.5. This is consistent with total war deaths being scale-free in the extreme tail of the distribution, which echos previous findings using the CoW conflict series.

Both the inverse Burr and log-normal models were fit to the entire data sample. As shown in Figure 6, the fits in the extreme tail deviate from the power-law fit in interesting ways. The inverse Burr under-predicts the probability of some of the most severe conflicts, even compared to the log-normal model. Meanwhile, it over predicts the likelihood of smaller conflicts relative to the log-normal. The fitted parameters for the inverse Burr are 611255.323, 0.669, and 0 for the two shape parameters and the scaling parameter respectively. The fitted parameters for the log-normal are 5.777 and 3.91 for the log-mean and log-standard deviation.

For global deaths per capita, the best-fitting power-law slope applies to all wars that resulted in at least 14 battle deaths (36.84% of observations) with a slope coefficient of 1.49. Again, this is consistent with war size being scale-free in the extreme tail of the distribution.

Table 1: GOF tests for total deaths

Model	GOF	p-value
Power-law	0.063	0.868
Inverse Burr	0.135	0.103
Log-normal	0.614	0.000

Based on 10,000 bootstraps.

As shown in the middle panel of Figure 6, the fits for the inverse Burr and log-normal in the extreme tail deviate less from the power-law fit compared to total deaths. However, the inverse Burr still under-predicts the probability of some of the most severe conflicts relative to the power-law, as does the log-normal. The fitted parameters for the inverse Burr are 752061.415, 0.567, and 0 for the two shape parameters and the scaling parameter respectively. The fitted parameters for the log-normal are 0.957 and 3.157 for the log-mean and log-standard deviation.

Finally, for belligerent deaths per capita, the best-fitting power-law slope applies to all wars that resulted in at least 222 battle deaths (84.21% of observations) with a slope coefficient of 1.62. Yet again, this is consistent with war size being scale-free in the extreme tail of the distribution.

As shown in the left panel of Figure 6, the fits for the inverse Burr and log-normal in the extreme tail deviate very little from the power-law fit. The inverse Burr slightly over-predicts the probability of some of the most severe conflicts relative to the power-law and the log-normal. The fitted parameters for the inverse Burr are 2.48, 0.614, and 21.102 for the two shape parameters and the scaling parameter respectively. The fitted parameters for the log-normal are 4.985 and 2.353 for the log-mean and log-standard deviation.

Goodness of Fit

Visual inspection of the models creates the impression that, while each fits much of the data relatively well, the quality of fit improves for the inverse Burr and log-normal when war size is normalized by population. This especially looks to be the case when war size is measured by belligerent deaths per capita. However, it would be unwise to simply select a model by *look*. This section discusses the results

Table 2: GOF tests for global deaths per capita

Model	GOF	p-value
Power-law	0.077	0.845
Inverse Burr	0.088	0.498
Log-normal	0.259	0.004

Based on 10,000 bootstraps.

from formal GOF tests that provide a more precise estimate for how well these models fit each of the measures.

Table 1 shows estimates from the simulation-based GOF test described previously in the paper for each of the models fit using total battle deaths. Recall that these tests infer goodness of fit by comparing how well a model fits the data assuming the model were the true data-generating process. For each of the models a KS distance statistic was calculated. Then multiple simulated datasets were generated using the estimated model parameters. The same model was then fit to each of the simulated datasets and the KS distance for that dataset calculated. A p-value was then computed by calculating the fraction of the simulations where the distance metric was greater than the one calculated with the original model estimates and data.

The results show that when conflict is measured by total deaths, the classic power-law model cannot be formally rejected: $D = 0.06$ (p-value = 0.87). The inverse Burr also cannot be rejected, however it only barely fails to cross the $p < 0.1$ threshold advised by Clauset, Shalizi, and Newman (2009). The log-normal model is the only one that can be rejected, with a $p \approx 0.000$.

The results are similar when we consider global deaths per capita. Table 2 reports the GOF test results for all the models when fit to this measure of war size. Both the classic power-law and inverse Burr models cannot be rejected while the log-normal can. The p-values associated with each of the tests are 0.85, 0.5, and 0 respectively. Notably, the inverse Burr makes for a much better fit for global deaths per capita than it did for total deaths. Where before the model escaped rejection by a fraction of a decimal point in its p-value, here it does so with much more breathing room.

The results differ yet again when considering belligerent deaths per capita. Table 3 reports GOF test

Table 3: GOF tests for belligerent deaths per capita

Model	GOF	p-value
Power-law	0.072	0.814
Inverse Burr	0.073	0.144
Log-normal	0.057	0.645

Based on 10,000 bootstraps.

results for each of the models fit using the intensity of war deaths per the populations of the countries fighting a war. Here, none of the models can be rejected. The p-values associated with each of the tests are 0.81, 0.14, and 0.65 respectively.

Model Comparisons

The results up to now show that the hypothesis that the data are power-law distributed in the extreme tail of the distribution cannot be rejected, regardless of whether war size is measured in terms of absolute severity or in relative terms (prevalence or intensity). However, alternatives also cannot be definitively rejected. The inverse Burr only narrowly escaped rejection for total battle deaths, while it more definitively failed to be rejected for global and belligerent deaths per capita. Further, while the log-normal could be rejected in the case of total deaths and global deaths per capita, it could not be rejected for belligerent deaths per capita. Clearly, since different alternatives are plausible fits for the data, additional tests are required to formally adjudicate which is best.

This section shows results from Vuong's (1989) likelihood ratio (LR) test, which is a non-nested model comparison test. As previously noted, a limitation of this approach is that it does not factor in the data truncation imposed by using the power-law model as compared to the alternatives when comparing model fit. However, at minimum, it provides a way to quantify how well the alternative models fit the data in the extreme tail of the the distribution.

Up first, Table 4 shows the results from Vuong's test applied to pairwise comparisons of the three competing models for total battle deaths. Recall that only the power-law and inverse Burr could not be rejected in the previous section, so from a data-driven perspective there already are grounds for

Table 4: Vuong's test for best fitting model for total battle deaths.

Models	Estimate	p-value
Power-law vs. Inverse Burr	0.854	0.393
Power-law vs. Log-normal	0.864	0.388
Inverse Burr vs. Log-normal	-2.384	0.017

Only non-truncated data points used for comparisons with the power-law. Full data used for inverse Burr vs. log-normal test.

selecting either of these. It is worth testing whether a model comparison test can provide additional guidance on which of these surviving models to select. However, according to Vuong's test the power-law is *not* statistically better than inverse Burr in the extreme tail of the distribution; albeit, the sign of the test statistic is positive, which means the likelihood for the power-law is greater. Interestingly, even though the log-normal could be rejected, it also does not perform statistically worse than the power-law in the extreme tail of the data. Even more interesting is the fact that the null of Vuong's test can be rejected when comparing the log-normal to the inverse Burr model, but the negative sign on the test indicates that the inverse Burr performs *worse* than the log-normal. This is unexpected because the log-normal model could be rejected in the model validation stage whereas the inverse Burr could not.

Table 5 shows results from Vuong's test when modeling global deaths per capita. In this case, none of the models can be rejected in favor of another. Recall again than in the previous GOF tests only the log-normal model could be rejected, meaning regardless of the results here, we do have a basis for narrowing down our options to either the power-law or inverse Burr. The signs on the test statistics indicate that the power-law does perform slightly better relative to the alternatives, but this difference is not statistically different from zero.

Finally, Table 6 shows results from Vuong's test applied to models fit to belligerent deaths per capita. Recall from the previous section that only in the case of war intensity was the null not rejected for each of the competing models. The failure to definitively reject at least one of the alternatives in the GOF tests makes the results from Vuong's test all the more vital in model selection. The estimates here seem to favor the log-normal model. In comparison to the power-law and inverse Burr, the log-normal's

Table 5: Vuong's test for best fitting model for global deaths per capita.

Models	Estimate	p-value
Power-law vs. Inverse Burr	0.064	0.949
Power-law vs. Log-normal	0.343	0.732
Inverse Burr vs. Log-normal	0.410	0.681

Only non-truncated data points used for comparisons with the power-law. Full data used for inverse Burr vs. log-normal test.

Table 6: Vuong's test for best fitting model for belligerent deaths per capita.

Models	Estimate	p-value
Power-law vs. Inverse Burr	0.328	0.743
Power-law vs. Log-normal	-0.153	0.878
Inverse Burr vs. Log-normal	-1.373	0.170

Only non-truncated data points used for comparisons with the power-law. Full data used for inverse Burr vs. log-normal test.

likelihood in the extreme tail of the distribution is higher. Unfortunately, as was the case in modeling battle death prevalence, this better performance is not statistically significant.

Implications

Evaluating the Long Peace

The foregoing analysis shows that the power-law and inverse Burr models survive the model validation tests when the variable of interest is total battle deaths or global deaths per capita. However, the inverse Burr only barely escapes rejection at the $p < 0.1$ level. Meanwhile, all three models (the power-law, inverse Burr, and log-normal) cannot be rejected for belligerent deaths per capita. Additional model comparison tests fail to provide a data-driven answer for which of the surviving alternatives should be preferred; though, the signs on the likelihood ratio tests suggest that the power-law may have an advantage over the others for total deaths and global deaths per capita and that the log-normal

may have an advantage over the inverse Burr and power-law for belligerent deaths per capita. Most puzzling, however, is the fact that the log-normal model is statistically better than the inverse Burr model for total deaths even though the inverse Burr was the one that passed muster in the model validation step while the log-normal did not.

It is tempting to stop here and simply conclude that peace scholars have some freedom of choice in studying the sizes of inter-state wars. While this is true, it would be a mistake to think this freedom of choice is free of consequence. Scholars should bear a few factors in mind when formal tests fail to offer a definitive answer about model selection.

The question that most occupies recent scholarship on war size is whether the so-called long peace is statistically detectable beginning in the mid-twentieth century. Many scholars argue that starting after World War II the international system entered an unprecedented period of peace, at least among major powers.[3] This argument is situated within a broader set of claims known as the decline-of-war thesis, which holds that wars over time, in addition to becoming less common, are becoming less deadly.

Some scholars answer in the affirmative on this question (Cunen, Hjort, and Nygård 2020; Pinker 2011; Spagat, Johnson, and Weezel 2018; Spagat and Weezel 2020) , while others answer in the negative (Braumoeller 2019; Clauset 2017, 2018). Most point to World War II as the relevant turning point, and Cunen, Hjort, and Nygård (2020) recently identified the year 1950 as the most statistically likely. This year was also identified as the most likely by Spagat and Weezel (2020).

One of the central points of tension in these studies is the choice of statistical model for identifying the long peace. As noted earlier, Cunen, Hjort, and Nygård (2020), who claim to recover evidence of the long peace, do so by using the inverse Burr distribution rather than the power-law. They also use total battle deaths as their measure of war size. However, as found in the previous section, the inverse Burr model is nearly rejected in a GOF test as a likely data-generating process for total battle deaths.[4]

[3] See Braumoeller (2019); Braumoeller (2021); and Cunen, Hjort, and Nygård (2020) for a comprehensive set of citations and summaries.

[4] As noted earlier, Cunen, Hjort, and Nygård (2020) use a fairly similar GOF test and get a p-value of 0.155 (see page 17 of the Online Appendix), meaning they fail to reject the inverse Burr. This obviously conflicts with the finding here, suggesting the need for more work to help explain why.

Moreover, the sign on the model comparison test favors the power-law over the inverse Burr in the extreme tail while both the sign and significance of the model comparison test favor the log-normal model.

But beyond the question of the appropriate statistical model of war size is the additional question of how model selection then influences whether evidence of the long peace can be recovered. If one model supports the long peace while another does not, this demonstrates all the more why a replicable data-driven approach to model selection is necessary. It may also show that when multiple models cannot be rejected, it may be unwise to choose only one to base results on.

To illustrate how choice of measure and model influences conclusions about the long peace, a bootstrapped test like that developed by Braumoeller (2019) is conducted for the power-law slopes fitted to pre- and post-1950 battle deaths using each of the measures of war size (total deaths, global deaths per capita, and belligerent deaths per capita). Though Braumoeller (2019) used 1945 as the cutoff in his analysis, the 1950 cutoff is used here since it was identified as the most probable change point by Cunen, Hjort, and Nygård (2020) and Spagat and Weezel (2020). Additional bootstrapped tests are performed using the inverse Burr and log-normal models.[5]

Tables 7, 8, and 9 report results from these statistical tests comparing pre-1950 to post-1950 trends in conflict size. The estimates are based on each of the alternative models of war size. Table 7 shows estimates from power-law fits for the data. Cell entries are the different power-law slopes, their difference, and the bootstrap p-value. The power-law slopes pre- and post-1950 are all less than 2, consistent with battle deaths being scale-free regardless of the time period and how conflict size is measured. Further, the power-law slopes in the different periods are not statistically different. That means we have little evidence that the most extreme wars before 1950 are generated by a different power-law distribution than wars after 1950.

Table 8 reports results using the inverse Burr model. As with the power-law, a bootstrap test is performed. Recall that the inverse Burr has three parameters. The analysis here homes in on the scaling

[5]Bootstrapping was done to maintain consistency with the bootstrapping procedure used for testing the equivalence of power-law slopes. However, well-behaved asymptotic inferences could just as easily been applied with the inverse Burr and log-normal models as well.

Table 7: A test of the long-peace using the classic power-law model.

Data	pre-1950	post-1950	Difference	p-value
Total Deaths	1.489	1.629	0.140	0.494
Global Deaths pc	1.535	1.641	0.106	0.415
Belligerent Deaths pc	1.673	1.667	-0.006	0.494

Entries are power-law slopes. 2,000 bootstraps performed.

Table 8: A test of the long-peace using the inverse Burr model.

Data	pre-1950	post-1950	Difference	p-value
Total Deaths	0.636	0.000102	-0.636	0.318
Global Deaths pc	3e-10	2.48e-07	2.48e-07	0.453
Belligerent Deaths pc	139	0.00177	-139	0.293

Entries are central tendency for inverse Burr. 2,000 bootstraps performed.

parameter μ which denotes the central tendency of the inverse Burr distribution. If this parameter is different between periods, this indicates a change in the expected rate of battle deaths. Cell entries in Table 8 are estimates of μ pre- and post-1950 along with their difference and the bootstrapped p-value. Like with the power-law, across alternative measures of battle deaths the inverse Burr central tendency pre- and post-1950 is not statistically different.[6]

Finally, Table 9 shows estimates from bootstrapped tests of the difference in the log-mean of battle deaths pre- and post-1950. As with the inverse Burr, the parameter μ (log-mean) captures the central tendency of the log-normal distribution. In the cases of total battle deaths and global deaths per capita, there is no statistically significant difference in war size pre- and post-1950. However, in the case of belligerent deaths per capita, there is. This finding is worth noting because only for belligerent deaths was the log-normal model not rejected as a plausible data-generating process for war size. In addition, though Vuong's test could not reject the null that the log-normal was superior to the power-law and inverse Burr, the signs of the test nonetheless favored the log-normal.

[6]This is not the same test that Cunen, Hjort, and Nygård (2020) use in their analysis. They instead apply a change-point algorithm that assumes a quasi-joint inverse Burr distribution before and after the optimal change-point. This may explain why their results are statistically significant whereas these are not. More research would need to be done to determine if their test is too lenient or if this bootstrapped test is too conservative.

Table 9: A test of the long-peace using the log-normal model.

Data	pre-1950	post-1950	Difference	p-value
Total Deaths	8.365	3.258	-5.107	0.257
Global Deaths pc	0.525	-1.623	-2.148	0.386
Belligerent Deaths pc	5.522	3.867	-1.654	0.005

Entries are central tendency for log-normal. 2,000 bootstraps performed.

The above results show that model selection and choice of measure can influence whether the long peace is statistically detectable. Importantly, only in the case of the log-normal model and when war size is measured as belligerent deaths per capita, is the long peace a statistically identifiable trend. This finding raises some important questions. On the one hand, it suggests that given enough models and options for how to quantify war size, it is possible to go on a successful fishing expedition for evidence of the long peace. On the other hand, this view may be too cynical, and it suggests (perhaps prematurely) that we should reject the one instance of significant results out of nine tests as a false positive. Focusing on this question misses the point, however. Rather, we should direct our attention to the fact that with researcher degrees of freedom comes a responsibility to engage in the model validation and selection exercises followed in this paper. This provides added transparency and a data-based justification for modeling choices that either support or refute important claims such as the long-peace.

Modeling Covariates

While the power-law is a justifiable choice across the measures of war size considered in this study, when either the inverse Burr or log-normal are supportable alternatives scholars should weigh their options carefully for an additional reason beyond the statistical implications discussed above. The power-law is not easily amenable to regression analysis. In fact, if the power-law slope is consistent with scale-free data (a slope less than 2), regression analysis is completely unjustifiable. This is not a limitation with the inverse Burr or log-normal models.

As Cunen, Hjort, and Nygård (2020) show, the inverse Burr model can be modified to accommodate

covariates. While an inverse Burr regression requires some additional programming to work since most software lack pre-packaged tools for making this happen, a log-normal regression is easy enough to estimate with standard statistical software, either via maximum likelihood or OLS using a log-transformation of the outcome. If researchers choose to go this route, they will find ample opportunity to make contributions to our understanding of war. While there are a couple exceptions (see Cunen, Hjort, and Nygård 2020; Weisiger 2013), almost no research has been done to assess correlates of war severity. Depending on how one wishes to operationalize war size, there is nothing preventing scholars from engaging in such an analysis.

Of course, the fact that the power-law is justifiable for wars in the extreme tail of the distribution across measures of war size should also be taken seriously. This study could not definitively rule out the idea that large wars are scale-free and thus prone to unpredictably large upswings in severity. Therefore, while better behaved statistical models are justifiable, the fact that such an extreme distribution may fit the data as well should encourage caution when applying tools like regression analysis to war size. Going forward, the best way to arrive at a more definitive conclusion about the best model for war size will be to collect more and better data on conflicts. As Clauset, Shalizi, and Newman (2009) show using simulations, when dealing with a sample of size $N \leq 100$, it is possible to fail to reject a model as the data-generating process for the data when the true data-generating process is some alternative. With an N of 95, the CoW dataset is subject to this very problem.

Data Truncation

A final implication that scholars should consider is data truncation. While the power-law remains a justifiable data-generating process for total battle deaths, global deaths per capita, and belligerent deaths per capita, this applies only for the most extreme wars in the data. For some of these measures upwards of 50% of the data is left unexplained. The inverse Burr and log-normal models provide a solution to this problem. If the goal is to maximize statistical precision and provide an explanation for the full set of wars in a conflict series, clearly the power-law is at a disadvantage.

However, while it is easy enough to quantify the extent of data truncation required to fit the

power-law to data, there is no straightforward way to quantify how this should factor in to model selection. It really comes down to the judgment of the researcher. Again, this speaks to the necessity of model validation and selection tests. These at least provide a replicable and transparent assessment of how competing models fit the data. From here, it may be advisable to either adopt a model on the basis of strong theoretical arguments or else remain agnostic about model choice and simply report results for each.

Conclusion

Research on the severity of wars is foundational to the quantitative study of conflict (e.g. Richardson 1948), but over time it has become a niche area of study as dyadic analyses of conflict onset have come to dominate much of the literature. This is beginning to change as questions about the severity of wars have recently received renewed attention both in the literature and in public discourse (e.g. Braumoeller 2019; Cunen, Hjort, and Nygård 2020; Pinker 2011). A central claim much of this research has attempted to test is the statistical significance of the long peace (a period defined by a decline in the frequency and severity of international war), purported to have begun in the second half of the twentieth century. The literature remains divided on whether the long peace is a statistically identifiable trend. Part of this division is driven by alternative data sources, but many studies rely on different statistical models of war size as well, often without a replicable or transparent justification for model selection.

This study makes a contribution to the conflict literature by using best-practices of model validation and selection proposed by Clauset, Shalizi, and Newman (2009) using the popular CoW dataset (Sarkees and Wayman 2010; D. J. Singer 1987) to see which among a set of alternative models truly best fit the data. The results show that, depending on how war size is measured, different models are plausible fits for war size.

Importantly, the specific findings of the analysis in this study should not be treated as definitive. Different conclusions may follow from different datasets or from studying different kinds of conflict,

such as civil wars. Furthermore, there are other models not considered here that might be just as applicable for studying war size. Finally, despite efforts to minimize mistakes in data collection, aggregation, and analysis in this study, errors may still have been made. In sum, scholars should support their modeling choices using their own data-driven assessments rather than those reported here or elsewhere.

It does seem clear, however, that conflict scholars need not limit themselves to using the power-law or any one other model to study the sizes of international conflicts. Multiple models may be justifiable. By using the methods outlined in here, researchers can find data-driven justification for alternative models that also grant greater latitude (such as the ability to conduct regression analysis) and statistical precision. This new-found freedom comes with consequences that should lead scholars to proceed with caution, however. Factors to consider include the fact that different models applied to the same measure of war size may lead to conflicting conclusions about trends such as the long peace. Another factor to consider is that some models justify regression analysis while others do not. In addition, some models require data truncation while others can be fit to a full dataset. Ultimately, if debate about trends such as the long peace is to proceed in a productive way, awareness of and transparency about these issues is necessary (but at the moment is in short supply).

Supplementary Code Appendix

```
# ==============================
# Setup, packages, and helpers
# ==============================

knitr::opts_chunk$set(cache=TRUE, echo=FALSE,
                      message=FALSE, warning=FALSE)

## Packages
library(tidyverse)
library(kableExtra)
library(poweRlaw)
library(actuar)
library(patchwork)
library(coolorrr)
theme_set(theme_test())
set_palette()

## Read in methods for inverse Burr to use with
## {poweRlaw} package. Code for the helper functions
## in the source file will be made available in a
## separate file since the code takes up a lot
## of space
source(
  here::here(
    "04_report",
    "00_coninvbur.R"
  )
)

# ====================================================
# Read in the data
# ====================================================
dt <- read_csv(
  here::here("01_data", "war-year.csv")
)

# ================================
# Analysis for section:
# "The Power Law and Alternatives"
# ================================
```

```r
## In-text percentages for Fig. 1
num <- dt$batdeath[dt$batdeath <= quantile(dt$batdeath, 0.8)]
den <- dt$batdeath
pct <- round(100 * sum(num) / sum(den), 2)

## Fig. 1 summarizing the density of total battle deaths
ggplot(dt) +
  aes(x = batdeath) +
  geom_density(fill = "gray") +
  scale_x_log10(
    labels = scales::comma
  ) +
  geom_vline(
    aes(
      xintercept = quantile(batdeath, 0.8)
    )
  ) +
  annotate(
    "text",
    x = quantile(dt$batdeath, 0.8),
    y = 0.35,
    label = "80th",
    hjust = 1
  ) +
  labs(
    x = "Battle Deaths (log-10)",
    y = "Density",
    fill = NULL
  ) +
  ggpal(aes = "fill")

## Fig. 2: example plots of inverse CDFs with power-law
x <- 1:100
y <- (x / max(x))^-4
p1 <- ggplot() +
  aes(x, y) +
  geom_line(size = 1) +
  labs(
    x = "x",
    y = "Pr(X > x)",
    title = "Unadjusted Scale"
  )
p2 <- p1 +
  scale_x_log10() +
```

```
  scale_y_log10() +
  labs(
    title = "log-log Scale"
  )
p1 + p2 &
  theme(
    axis.text = element_blank()
  )

## Fig. 3: the empirical inverse CDF shown for inensity of battle deaths
f <- function(x) rank(-x) / max(rank(-x))
ggplot(dt) +
  aes(x = batdeathpc * 1000000, y =f(batdeathpc)) +
  geom_point(color = "gray") +
  scale_x_log10(
    labels = scales::comma
  ) +
  scale_y_log10() +
  labs(
    x = "Battle Deaths per Million",
    y = "Pr(X > x)"
  )

## Fig. 4: Example inverse Burr inverse CDF
x <- 1:100
y <- (1 / (1 + exp(x)))^2
p1 <- ggplot() +
  aes(x, y) +
  geom_line(size = 1) +
  labs(
    x = "x",
    y = "Pr(X > x)",
    title = "Unadjusted Scale"
  )
p2 <- p1 +
  scale_x_log10() +
  scale_y_log10() +
  labs(
    title = "log-log Scale"
  )
p1 + p2 &
  theme(
    axis.text = element_blank()
  )
```

```r
## Fig. 5: Example log-normal inverse CDF
x <- 1:100
y <- 1 - pnorm(x, mean = 0, sd = 20)
p1 <- ggplot() +
  aes(x, y) +
  geom_line(size = 1) +
  labs(
    x = "x",
    y = "Pr(X > x)",
    title = "Unadjusted Scale"
  )
p2 <- p1 +
  scale_x_log10() +
  scale_y_log10() +
  labs(
    title = "log-log Scale"
  )
p1 + p2 &
  theme(
    axis.text = element_blank()
  )

# ================================
# Emperical analysis for section:
# "Analysis: Model Fitting"
# ================================

# the data
x <- dt$batdeath
y <- dt$batdeath / dt$wpop * 1000000
z <- dt$batdeathpc * 1000000

# fits for total war size
x1 <- conpl$new(x)
x1$setXmin(estimate_xmin(x1, xmax = 1e09))
x2 <- coninvburr$new(x)
x2$setPars(estimate_pars(x2))
x3 <- conlnorm$new(x)
x3$setPars(estimate_pars(x3))
x4 <- dispois$new(x)
x4$setPars(estimate_pars(x4))

# fits for "all cause mortality"
y1 <- conpl$new(y)
```

```
y1$setXmin(estimate_xmin(y1, xmax = 1e09))
y2 <- coninvburr$new(y)
y2$setPars(estimate_pars(y2))
y3 <- conlnorm$new(y)
y3$setPars(estimate_pars(y3))
y4 <- dispois$new(ceiling(y))
y4$setPars(estimate_pars(y4))

# fits for risk of war
z1 <- conpl$new(z)
z1$setXmin(estimate_xmin(z1, xmax = 1e09))
z2 <- coninvburr$new(z)
z2$setPars(estimate_pars(z2))
z3 <- conlnorm$new(z)
z3$setPars(estimate_pars(z3))
z4 <- dispois$new(ceiling(z))
z4$setPars(estimate_pars(z4))

# Fig. 7: plot the results
par(mfcol = c(1, 3))
plot(x1, pch = 19, col = "gray",
     xlab = "War Size", ylab = "Pr(X > x)",
     main = "Total Deaths")
lines(x3, col = "red3", lwd = 3)
lines(x2, col = "steelblue", lwd = 3)
lines(x1, col = "black", lwd = 3)
plot(y1, pch = 19, col = "gray",
     xlab = "War Size", ylab = "Pr(X > x)",
     main = "Global Deaths per Capita")
lines(y3, col = "red3", lwd = 3)
lines(y2, col = "steelblue", lwd = 3)
lines(y1, col = "black", lwd = 3)
plot(z1, pch = 19, col = "gray",
     xlab = "War Size", ylab = "Pr(X > x)",
     main = "Belligerent Deaths per Capita")
lines(z3, col = "red3", lwd = 3)
lines(z2, col = "steelblue", lwd = 3)
lines(z1, col = "black", lwd = 3)
legend(
  "bottomleft",
  lty = c(1, 1, 1),
  col = c("black", "steelblue", "red3"),
  legend = c("Power-law", "Inverse Burr", "Log-normal"),
  bty = "n",
```

```
  lwd = c(3, 3, 3)
)

# ================================
# Emperical analysis for section:
# "Analysis: Goodness of Fit"
# ================================

# GOF tests for battle death *severity* using 10,000 bootstraps
set.seed(1)
gof1 <- my_bootstrap_p(
  x1,
  threads = 4, no_of_sims = 10000,
  xmins = rep(x1$xmin, 2),
  seed = 1
)
gof2 <- my_bootstrap_p(
  x2, no_of_sims = 10000,
  seed = 1
)
gof3 <- my_bootstrap_p(
  x3, no_of_sims = 10000,
  threads = 4,
  xmins = rep(min(x), 2),
  seed = 1
)

tibble( # report in a table (Table 1)
  Model = c("Power-law", "Inverse Burr", "Log-normal"),
  GOF = c(gof1$gof, gof2$gof, gof3$gof),
  "p-value" = c(gof1$p, gof2$p, gof3$p)
) |>
  kbl(
    caption = "GOF tests for total deaths",
    booktabs = T,
    linesep = "",
    digits = 3
  ) |>
  add_footnote(
    "Based on 10,000 bootstraps.",
    notation = "none"
  )
```

```r
# GOF tests for battle death *prevalence* with 2,000 bootstraps
set.seed(1)
gof1 <- my_bootstrap_p(
  y1,
  threads = 4, no_of_sims = 10000,
  xmins = rep(y1$xmin),
  seed = 1
)
gof2 <- my_bootstrap_p(
  y2, no_of_sims = 10000,
  seed = 1
)
gof3 <- my_bootstrap_p(
  y3,
  threads = 4,
  xmins = rep(min(y), 2),
  no_of_sims = 10000,
  seed = 1
)

tibble( # report in a table (Table 2)
  Model = c("Power-law", "Inverse Burr", "Log-normal"),
  GOF = c(gof1$gof, gof2$gof, gof3$gof),
  "p-value" = c(gof1$p, gof2$p, gof3$p)
) |>
  kbl(
    caption = "GOF tests for global deaths per capita",
    booktabs = T,
    linesep = "",
    digits = 3
  ) |>
  add_footnote(
    "Based on 10,000 bootstraps.",
    notation = "none"
  )


# GOF tests for battle death *intensity* with 2,000 bootstraps
set.seed(1)
gof1 <- my_bootstrap_p(
  z1,
  threads = 4,
  no_of_sims = 10000,
  xmins = rep(z1$xmin, 2),
```

```
  seed = 1
)
gof2 <- my_bootstrap_p(
  z2, no_of_sims = 10000,
  seed = 1
)
gof3 <- my_bootstrap_p(
  z3,
  threads = 4,
  no_of_sims = 10000,
  xmins = rep(min(z), 2),
  seed = 1
)

tibble( # report in table (Table 3)
  Model = c("Power-law", "Inverse Burr", "Log-normal"),
  GOF = c(gof1$gof, gof2$gof, gof3$gof),
  "p-value" = c(gof1$p, gof2$p, gof3$p)
) |>
  kbl(
    caption = "GOF tests for belligerent deaths per capita",
    booktabs = T,
    linesep = "",
    digits = 3
  ) |>
  add_footnote(
    "Based on 10,000 bootstraps.",
    notation = "none"
  )


# ================================
# Emperical analysis for section:
# "Analysis: Model Comparisons"
# ================================

# A function to compare model likelihoods
# only for all X >= xmin they have in common.
simp_compare <- function(d1, d2) {
  if(d1$xmin == d2$xmin) {
    cp <- compare_distributions(d1, d2)
  } else {
    d1cpy <- d1$copy()
    d2cpy <- d2$copy()
```

```
    d1cpy$setXmin(max(d1$xmin, d2$xmin))
    d2cpy$setXmin(max(d1$xmin, d2$xmin))
    cp <- compare_distributions(d1cpy, d2cpy)
  }
  tibble(
    Models = paste0(
      class(d1)[1], " vs. ",
      class(d2)[1]
    ),
    Estimate = cp$test_statistic,
    "p-value" = cp$p_two_sided
  )
}

# perform model comparisons for battle death *severity* and
# report in a table (Table 4)
bind_rows(
  simp_compare(x1, x2),
  simp_compare(x1, x3),
  simp_compare(x2, x3)
) |>
  mutate(
    Models = c(
      "Power-law vs. Inverse Burr",
      "Power-law vs. Log-normal",
      "Inverse Burr vs. Log-normal"
    )
  ) |>
  kbl(
    caption = "Vuong's test for best fitting model for total battle deaths.",
    booktabs = T,
    digits = 3,
    linesep = ""
  ) |>
  add_footnote(
    "Only non-truncated data points used for comparisons with the power-law. Full data u
    notation = "none"
  )

# perform model comparisons for battle death *prevalence* and
# report in a table (Table 5)
bind_rows(
  simp_compare(y1, y2),
  simp_compare(y1, y3),
```

```
  simp_compare(y2, y3)
) |>
  mutate(
    Models = c(
      "Power-law vs. Inverse Burr",
      "Power-law vs. Log-normal",
      "Inverse Burr vs. Log-normal"
    )
  ) |>
  kbl(
    caption = "Vuong's test for best fitting model for global deaths per capita.",
    booktabs = T,
    digits = 3,
    linesep = ""
  ) |>
  add_footnote(
    "Only non-truncated data points used for comparisons with the power-law. Full data u
    notation = "none"
  )

# perform model comparisons for battle death *intensity* and
# report in a table (Table 6)
bind_rows(
  simp_compare(z1, z2),
  simp_compare(z1, z3),
  simp_compare(z2, z3)
) |>
  mutate(
    Models = c(
      "Power-law vs. Inverse Burr",
      "Power-law vs. Log-normal",
      "Inverse Burr vs. Log-normal"
    )
  ) |>
  kbl(
    caption = "Vuong's test for best fitting model for belligerent deaths per capita.",
    booktabs = T,
    digits = 3,
    linesep = ""
  ) |>
  add_footnote(
    "Only non-truncated data points used for comparisons with the power-law. Full data u
    notation = "none"
  )
```

```r
# ============================================
# Empirical analysis for section:
# "Implications: Identifying the 'Long-Peace'"
# ============================================

# set up for empirical test of the long peace
library(furrr) # use parallel computing
plan(multicore, sessions = 7) # 7 cores on my Intel i7 machine

# a wrapper for fitting and boostrapping the power-law
pl_fit <- function(dat, its = 2000) {
  # fit the model
  m <- conpl$new(dat)
  m$setXmin(estimate_xmin(m, xmax = 1e09))
  # perform bootstrap
  tibble(
    it = 1:its,
    bm = future_map(
      it, ~ {
        sdat <- sample(dat, length(dat), T)
        bm <- conpl$new(sdat)
        bm$setXmin(estimate_xmin(bm, xmax = 1e09))
        tibble(par = bm$pars)
      },
      .options = furrr_options(seed = T)
    )
  ) -> boot_out
  # return fit and bootstrap
  list(
    pars = m$pars,
    boot_pars = boot_out |>
      unnest(bm)
  )
}

# a wrapper for fitting and bootstrapping the inverse Burr
ib_fit <- function(dat, its = 2000) {
  m <- coninvburr$new(dat)
  m$setPars(estimate_pars(m))
  tibble(
    it = 1:its,
    bm = future_map(
      it, ~ {
        sdat <- sample(dat, length(dat), T)
```

```
        bm <- coninvburr$new(sdat)
        bm$setPars(estimate_pars(bm))
        tibble(
          par1 = bm$pars[1],
          par2 = bm$pars[2],
          par3 = bm$pars[3]
        )
      },
      .options = furrr_options(seed = T)
    )
  ) -> boot_out
  list(
    pars = m$pars,
    boot_pars = boot_out |>
      unnest(bm)
  )
}

# a wrapper for fitting and bootstrapping the log-normal
ln_fit <- function(dat, its = 2000) {
  m <- conlnorm$new(dat)
  m$setPars(estimate_pars(m))
  tibble(
    it = 1:its,
    bm = future_map(
      it, ~ {
        sdat <- sample(dat, length(dat), T)
        bm <- conlnorm$new(sdat)
        bm$setPars(estimate_pars(bm))
        tibble(
          par1 = bm$pars[1],
          par2 = bm$pars[2]
        )
      },
      .options = furrr_options(seed = T)
    )
  ) -> boot_out
  list(
    pars = m$pars,
    boot_pars = boot_out |>
      unnest(bm)
  )
}
```

```r
# a quick function to compute p-values from bootstraps
get_p <- function(x, y) {
  2 * min(
    mean(x > y),
    mean(x < y)
  )
}

# divide the data by pre- and post-1950
  # battle death *severity*
xpre <- x[dt$year <= 1950]
xpos <- x[dt$year > 1950]

  # battle death *prevalence*
ypre <- y[dt$year <= 1950]
ypos <- y[dt$year > 1950]

  # battle death *intensity*
zpre <- z[dt$year <= 1950]
zpos <- z[dt$year > 1950]

# power-law fits
set.seed(1)
x1pre <- pl_fit(xpre)
x1pos <- pl_fit(xpos)
y1pre <- pl_fit(ypre)
y1pos <- pl_fit(ypos)
z1pre <- pl_fit(zpre)
z1pos <- pl_fit(zpos)

# summarize results
pl_tests <- tibble(
  Data = c("Total Deaths", "Global Deaths pc", "Belligerent Deaths pc"),
  "pre-1950" = c(x1pre$pars, y1pre$pars, z1pre$pars),
  "post-1950" = c(x1pos$pars, y1pos$pars, z1pos$pars),
  Difference = `post-1950` - `pre-1950`,
  "p-value" = c(
    get_p(x1pre$boot_pars$par, x1pos$boot_pars$par),
    get_p(y1pre$boot_pars$par, y1pos$boot_pars$par),
    get_p(z1pre$boot_pars$par, z1pos$boot_pars$par)
  )
)

# report in a table (Table 7)
```

```
kbl(
  pl_tests,
  caption = "A test of the long-peace using the classic power-law model.",
  digits = 3,
  booktabs = T,
  linesep = ""
) |>
  add_footnote(
    "Entries are power-law slopes. 2,000 bootstraps performed.",
    notation = "none"
  )

# inverse Burr fits
x2pre <- ib_fit(xpre)
x2pos <- ib_fit(xpos)
y2pre <- ib_fit(ypre)
y2pos <- ib_fit(ypos)
z2pre <- ib_fit(zpre)
z2pos <- ib_fit(zpos)

# summarize results
ib_tests <- tibble(
  Data = c("Total Deaths", "Global Deaths pc", "Belligerent Deaths pc"),
  "pre-1950" = c(x2pre$pars[3], y2pre$pars[3], z2pre$pars[3]),
  "post-1950" = c(x2pos$pars[3], y2pos$pars[3], z2pos$pars[3]),
  Difference = `post-1950` - `pre-1950`,
  "p-value" = c(
    get_p(x2pre$boot_pars$par3, x2pos$boot_pars$par3),
    get_p(y2pre$boot_pars$par3, y2pos$boot_pars$par3),
    get_p(z2pre$boot_pars$par3, z2pos$boot_pars$par3)
  )
)

# report in a table (Table 8)
kbl(
  ib_tests |> mutate(
    across(
      2:4, ~ signif(.x, 3) |> as.character()
    )
  ),
  caption = "A test of the long-peace using the inverse Burr model.",
  digits = 3,
  booktabs = T,
  linesep = ""
```

```
) |>
  add_footnote(
    "Entries are central tendency for inverse Burr. 2,000 bootstraps performed.",
    notation = "none"
  )

# log-normal fits
x3pre <- ln_fit(xpre)
x3pos <- ln_fit(xpos)
y3pre <- ln_fit(ypre)
y3pos <- ln_fit(ypos)
z3pre <- ln_fit(zpre)
z3pos <- ln_fit(zpos)

# summarize results
ln_tests <- tibble(
  Data = c("Total Deaths", "Global Deaths pc", "Belligerent Deaths pc"),
  "pre-1950" = c(x3pre$pars[1], y3pre$pars[1], z3pre$pars[1]),
  "post-1950" = c(x3pos$pars[1], y3pos$pars[1], z3pos$pars[1]),
  Difference = `post-1950` - `pre-1950`,
  "p-value" = c(
    get_p(x3pre$boot_pars$par1, x3pos$boot_pars$par1),
    get_p(y3pre$boot_pars$par1, y3pos$boot_pars$par1),
    get_p(z3pre$boot_pars$par1, z3pos$boot_pars$par1)
  )
)

# report in a table (Table 9)
kbl(
  ln_tests,
  caption = "A test of the long-peace using the log-normal model.",
  digits = 3,
  booktabs = T,
  linesep = ""
) |>
  add_footnote(
    "Entries are central tendency for log-normal. 2,000 bootstraps performed.",
    notation = "none"
  )
```

References

Braumoeller, Bear F. 2019. *Only the Dead: The Persistence of War in the Modern Age*. New York: Oxford University Press.

———. 2021. "Trends in Interstate Conflict." In *What Do We Know about War? Third Edition*, edited by Sara McLaughlin Mitchell and John A. Vasquez. Rowman; Littlefield.

Cederman, Lars-Erik. 2003. "Modeling the Size of Wars: From Billiard Balls to Sandpiles." *American Political Science Review* 97 (1): 135–59.

Cederman, Lars-Erik, T. Camber Warren, and Didier Sornette. 2011. "Testing Clausewitz: Nationalism, Mass Mobilization, and the Severity of War." *International Organization* 65 (4): 605–38.

Cirillo, Pasquale, and Nassim Nicholas Taleb. 2016. "On the Statistical Properties and Tail Risk of Violent Conflicts." *Physica A: Statistical Mechanics and Its Applications* 452: 29–45.

Clauset, Aaron. 2017. "The Enduring Threat of a Large Interstate War." Technical report. One Earth Foundation.

———. 2018. "Trends and Fluctuations in the Severity of Interstate Wars." *Science Advances* 4 (2): eaao3580.

Clauset, Aaron, Cosma Rohilla Shalizi, and M.E.J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51 (4): 661–703.

Cunen, Céline, Nils Lid Hjort, and Håvard Mokleiv Nygård. 2020. "Statistical Sightings of Better Angels: Analysing the Distribution of Battle-Deaths in Interstate Conflict over Time." *Journal of Peace Research* 57 (2): 221–34.

Miller, Steven V. 2022. "peacesciencer: An r Package for Quantitative Peace Science Research." *Conflict Management and Peace Science*. http://svmiller.com/peacesciencer/.

Obermeier, Anna Marie, and Siri Aas Rustad. 2023. "Conflict Trends: A Global Overview, 1946–2022." *Oslo: Peace Research Institute Oslo (PRIO). November* 1 (2023): 201946–2022.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York: Viking Adult.

Reiter, Dan, Allan C Stam, and Michael C Horowitz. 2016. "A Deeper Look at Interstate War Data: Interstate War Data Version 1.1." *Research & Politics* 3 (4): 2053168016683840.

Richardson, Lewis F. 1948. "Variation of the Frequency of Fatal Quarels with Magnitude." *American Statistical Association* 43: 523–46.

———. 1960. *Statistics of Deadly Quarels*. Chicago: Quadrangle Books.

Sarkees, Meredith Reid, and Frank Wayman. 2010. *Resort to War: 1816 - 2007*. Washington DC: CQ Press.

Singer, David J. 1987. "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816-1985." *International Interactions* 14 (1): 115–32.

Singer, J. David, Stuart A. Bremer, and John Stuckey. 1972. "Capability Distribution, Uncertainty, and Major Power War, 1820-1965." In *Peace, War and Numbers*, edited by Bruce Russett. Beverly Hills, CA: Sage Publications, Inc.

Spagat, Michael, Neil F Johnson, and Stijn van Weezel. 2018. "Fundamental Patterns and Predictions of Event Size Distributions in Modern Wars and Terrorist Campaigns." *PloS One* 13 (10): e0204639.

Spagat, Michael, and Stijn van Weezel. 2020. "The Decline of War Since 1950: New Evidence." In *Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences*, edited by Nils Peter Gleditsch, 129–42. Springer.

Taleb, Nassim Nicholas. 2010. *The Black Swan: The Impact of the Highly Improbable*. New York: Random House.

Verbeeck, Cis, Emil Kraaikamp, Daniel F. Ryan, and Olena Podladchikova. 2019. "Solar Flare Distributions: Lognormal Instead of Power Law?" *The Astrophysical Journal* 884 (50): 1–16.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica: Journal of the Econometric Society*, 307–33.

Weisiger, Alex. 2013. *Logics of War: Explanations for Limited and Unlimited Conflicts*. Cornell University Press.